Pipelines de ETL e Machine Learning com Apache Spark

Serving de Modelos de Machine Learning

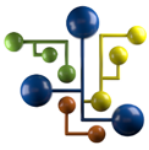
O Serving de Modelos de Machine Learning é o processo de disponibilizar um modelo treinado para uso em produção, permitindo que ele receba novos dados e retorne previsões em tempo real ou em lote. É uma etapa fundamental no ciclo de vida de um projeto de Machine Learning, pois conecta o modelo ao mundo real, integrando-o a sistemas ou aplicações que dependem de suas previsões para tomadas de decisão.

No contexto de serving, o modelo é geralmente implantado como um serviço acessível via API (como REST ou gRPC). Quando uma aplicação cliente envia uma solicitação contendo dados de entrada, o modelo processa esses dados e retorna a previsão correspondente. Essa infraestrutura precisa ser confiável, escalável e capaz de lidar com diferentes demandas, desde poucas requisições ocasionais até alto volume de chamadas em sistemas críticos.

O serving também pode incluir funcionalidades adicionais, como:

- Gerenciamento de versões: Para acompanhar atualizações ou substituições do modelo.
- Monitoramento: Para rastrear a qualidade das previsões e identificar problemas, como degradação do modelo devido a mudanças nos dados (data drift).
- Recursos de baixo tempo de resposta: Para aplicações que requerem previsões quase instantâneas.
- Acompanhamento de logs: Para análise de desempenho e auditoria.

Em plataformas como Databricks, AWS SageMaker ou Google AI Platform, o serving de modelos é oferecido como um serviço gerenciado, facilitando a implantação e integração. Em projetos personalizados, o modelo pode ser servido usando bibliotecas e frameworks como Flask, FastAPI ou TensorFlow Serving.



Equipe DSA

Muito Obrigado!
Continue Trilhando Uma Excelente Jornada de Aprendizagem.